Fusion of machine learning algorithms for building prognostic models in non-small cell lung cancer using clinical and radiomics features from ^{18}F-FDG PET/CT images

S. SEPEHRI[1], T. UPADHAYA[2], D. VISVIKIS[1], C. CHEZE LE REST[1, 2], M. HATT[1]

[1] LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France
[2] Département de médecine nucléaire, CHU Milétrie, Poitiers, France

[1]Shima.Sepehri@univ-brest.fr, hatt@univ-brest.fr, Dimitris@univ-brest.fr
[2] Taman.Upadhaya@chu-poitiers.fr, Catherine.Cheze-Le-Rest@chu-poitiers.fr

Funding: This work has received support from the DGOS and INCa through the project PRINCE (PRTK 2015)

Abstract – Radiomics is the high-throughput extraction of quantitative features from medical images, used to build prognostic and predictive models for personalized medicine. Given the large number of extracted features with respect to the small number of patients' data available, machine (deep) learning (ML) has become a major component of radiomics analyses. In this work we compared three ML methods, namely random forest (RF), support vector machines (SVM), both with embedded features selection, and logistic regression (LR) with stepwise feature selection, for building prognostic models exploiting clinical and ^{18}F-FDG PET radiomics features in lung cancer patients. Moreover, we are interested in determining if the fusion of these methods outputs could improve the final prediction. Our results show that the different ML pipelines select different sets of features and reach different classification performance, with a relatively moderate agreement, which is why the fusion of their outputs can help reach a higher performance (accuracy of 71% for the fusion using majority voting, compared to 67, 64 and 63% for RF, SVM and LR respectively). Even though the level of accuracy reached can seem relatively low (~70%), the resulting prognostic stratification is higher than when relying on clinical stage (61%), and of interest for clinical practice as it could help identifying patients with higher risk amongst stage II and III patients, that could benefit from intensified treatment and/or more frequent follow-up after treatment.

1 Introduction

Lung cancer is still a deadly disease, despite improvements in diagnosis, staging and treatment. It remains the first cause of cancer death for men and the second for women [1]. The variability in outcomes remain vast, with significant differences between patients depending on the stage of the disease. Staging is indeed one of the major clinical criteria on which physicians rely in order to choose a therapeutic strategy (*i.e.*, concomitant or sequential combination of surgery, chemotherapy and radiotherapy) [2]. However, even amongst patients with a similar disease stage, especially for stage II and III, there can be highly variable outcomes (*i.e.*, response to therapy and survival). ^{18}F-FDG Positron Emission Tomography / Computed Tomography (PET/CT) (figure 1) is the standard medical imaging modality for lung cancer diagnosis, staging and treatment planning and monitoring [3].

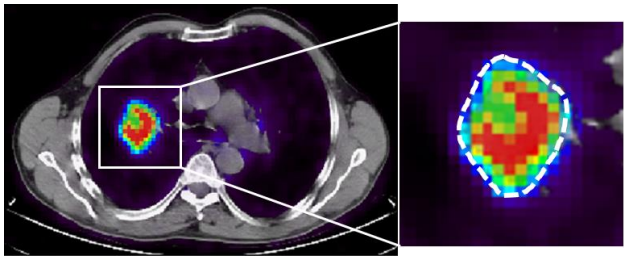

Figure 1: left: an FDG PET/CT image of a lung cancer patient (axial slice of a 3D volume acquisition at the tumor level in the lungs). PET and CT datasets are fused for visualization: CT appears in grey levels while PET appears in false colors, red indicating higher uptake of FDG radiotracer and thus higher tumor metabolism, quantitatively measured as standardized uptake values (SUV). Right: zoom-in of tumor, showing the automated contour (dashed white) to define the tumor volume, in which handcrafted radiomics features (shape, intensity, textures) are calculated.

However, these highly informative quantitative medical images are still used routinely in clinical practice only through visual examination and for diagnosis and staging only. Radiomics consists in the extraction of a large amount of handcrafted quantitative features from medical images that are subsequently processed by machine learning algorithms in order to build models (combining these radiomics features with other clinical variables such as gender or stage) to identify a tumor type, correlate with underlying biological information, or predict outcomes [4].

In the present work we focus on the comparison of different machine learning techniques as well as their fusion, for the goal of building models predictive of outcome (prognosis) in lung cancer, based on radiomics features extracted from the FDG PET component of the PET/CT images.

2 Material and methods

2.1 Patients data

A cohort of 138 non-small cell lung cancer (NSCLC) patients with stage 2 and 3 was retrospectively (n=87) and prospectively (n=51) collected at the CHU Milétrie in Poitiers, France. Treatment was mostly (chemo) radiotherapy (surgery mostly concerns stage I patients, whereas metastatic patients with stage IV disease have a whole different management). All patients had a ^{18}F-FDG PET/CT acquired as part of their diagnosis and staging procedure prior to treatment, which was exploited in the present work. The cohort was split into a training set (67%, n=92) and a testing set (33%, n=46) using stratified sampling, ensuring similar outcome, number of events and clinical characteristics in both sets. The set endpoint for this study was to identify patients with poor overall survival (OS). Median OS in the present cohort was 14.7 months.

The local ethics committee board approved this study.

2.2 Radiomics and machine learning workflow

Regarding the radiomics workflow, primary tumor volumes were characterized after automated segmentation of the PET image using the Fuzzy Locally Adaptive Bayesian (FLAB) algorithm [5], [6]. The handcrafted radiomics features were intensity metrics (n=10), shape descriptors (n=14) and textural features (n=66). They were extracted from the 3D delineated tumor volumes in PET and were validated with the most up-to-date international Image Biomarker Standardization Initiative (IBSI) reference definition and benchmark [7], [8]. Regarding the 2nd- and higher-order textural features, three different grey-level discretization methods, namely histogram equalization, fixed number of bins (using 64 bins) and fixed bin-width interval (bin width of 0.5 SUV) were considered. Texture matrices were implemented in 3D with the merging strategy, considering all directions simultaneously. A total of 223 features were thus extracted from each PET tumor volume for each patient.

All available clinical variables (gender, stage, smoking history, histology, treatment modality, etc.) and radiomics features were entered in the three different ML pipelines (features selection and classifier) under comparison, namely Support Vector Machines (SVM) with Recursive Feature Elimination (RFE), Random Forests (RF) with Embedded Wrapper (EW) method, and Logistic Regression (LR) with the stepwise method. The classification task was defined as to identify for each patient whether its OS would be $>$ or $\leq$ median OS. The performance of this binary and balanced (since the aim is to classify above or below median survival) classification was evaluated using accuracy, sensitivity and specificity. For each method, the best model was chosen in the training set according to the following criteria: number of required features (for similar levels of accuracy, a lower number of features is preferred for reducing the risk of overfitting and for higher chance of generalizability and better performance in external testing sets), accuracy and balance between sensitivity and specificity. Finally, in order to generate a consensus of the outputs from the

three ML pipelines, we implemented a simple majority-voting rule.

3 Results

In the training set, the best model built by RF with a reasonable number of features (25) reached an accuracy of 89%. In the testing set, this model obtained 67% accuracy. With SVM, the best model combining a small number of features (27) obtained an accuracy of 100%, however in the testing set, this model obtained only 64% accuracy. Higher accuracy (69%) could be obtained, at the cost of including a much larger number of features, which would likely reduce the likelihood the model could perform well in external datasets. The logistic regression reached 72% accuracy in training relying on 37 features, with 63% accuracy in testing.

The agreement between the three ML pipelines was moderate suggesting potential improvement could be provided by fusing the outputs. The fusion of the three methods with majority voting indeed led to an increased accuracy of 71% in the testing set (Tab 1.), demonstrating a slightly higher performance than each of the ML pipelines independently. By comparison, the accuracy reached by the standard clinical factor routinely used to stratify patients and determine therapeutic options (stage 2 vs. 3), was 61% and 58% in the training and testing sets respectively.

Tab 1: Comparison results of ML pipelines

ML Method	Training Accuracy	Testing Accuracy	# Features
RF	89	67	25
SVM	100	64	27
LR	72	63	37
Fusion (majority voting)	1	71	

4 Conclusion

Our study has a few limitations. It is a mostly retrospective and monocentric study. Only primary tumor volumes, not involved lymph nodes, were analyzed. We did not include features extracted from the low-dose CT component. We included and compared only three ML pipelines, however we selected optimized ones relying on embedded feature selection approaches.

Nonetheless, our results obtained in a rather large group of patients can be of interest for the community as they show that even optimized ML pipelines select different sets of features and reach different classification performance, with a relatively moderate agreement, which is why the fusion of their outputs can help reaching a higher performance (accuracy of 71% for the fusion using majority voting, compared to 67, 64 and 63% for RF, SVM and LR respectively). Even though the level of accuracy reached can seem relatively low compared to other applications, it is important to emphasize that prognostic prediction in lung cancer is notoriously difficult and that even such a moderate accuracy actually leads to a higher prognostic stratification than clinical staging, and is therefore of real interest for clinical practice, as it could help identifying patients with higher risk amongst stage II and III patients. These higher risk patients could benefit from intensified treatment and/or more frequent follow-up after treatment.

Future work will consist in adding the radiomics features from the low-dose CT component and considering other ML methods and fusion approaches. Comparison with deep learning techniques is also ongoing. Finally, validation of the developed models in external cohorts will be carried out as well.

5 References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] E. Coche, M. Lonneux, and X. Geets, "Lung cancer: Morphological and functional approach to screening, staging and treatment planning," *Future Oncol*, vol. 6, no. 3, pp. 367–80, Mar. 2010.

[3] A. W. Sauter, N. Schwenzer, M. R. Divine, B. J. Pichler, and C. Pfannenberg, "Image-derived biomarkers and multimodal imaging strategies for lung cancer management," *Eur. J. Nucl. Med. Mol. Imaging*, Jan. 2015.

[4] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: extracting more information from medical

images using advanced feature analysis," *Eur J Cancer*, vol. 48, no. 4, pp. 441–6, Mar. 2012.

[5] M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," *IEEE Trans Med Imaging*, vol. 28, no. 6, pp. 881–93, Jun. 2009.

[6] M. Hatt, C. Cheze le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, and D. Visvikis, "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications," *Int J Radiat Oncol Biol Phys*, vol. 77, no. 1, pp. 301–8, May 2010.

[7] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, and for the I. B. S. Initiative, "Image biomarker standardisation initiative," *ArXiv161207003 Cs*, Dec. 2016.

[8] M. Hatt, M. Vallieres, D. Visvikis, and A. Zwanenburg, "IBSI: an international community radiomics standardization initiative," *J. Nucl. Med.*, vol. 59, no. supplement 1, pp. 287–287, Jan. 2018.